How do auditory and visual instruction modalities in an AR escape room affect user task performance and user experience?

Andreas Garcia
Kungliga Tekniska Högskolan
Stockholm, Sweden
agarc@kth.se

Alexander Kazakov
Kungliga Tekniska Högskolan
Stockholm, Sweden
akaz@kth.se

Louisa Helen Hirvonen
Kungliga Tekniska Högskolan
Stockholm, Sweden
lhhi@kth.se

Matilda Jansson
Kungliga Tekniska Högskolan
Stockholm, Sweden
matilja@kth.se

Abstract

This report details the design, implementation and evaluation of an augmented reality (AR) escape room. AR offers a way to realise immersive, engaging experiences. The escape room was built in Unity and deployed on an iPad and combines visual, auditory and haptic modalities. Players must find and complete three puzzles to escape the room, represented by three different animals: Luna the Cat, Peri the Parrot, and Kaa the Cobra. Building upon existing research on the effect of AR instruction modalities in other domains, two modes of instruction were compared: auditory instruction in the form of speech, and visual instruction in the form of text. There was a significant difference in task performance, with users in the auditory instruction group completing the escape room significantly faster. However, user experience, measured by the spatial interaction (SPINE) questionnaire, did not significantly differ between groups except for in terms of navigation. Overall, the results indicate that the relationship between instruction modalities, task performance and user experience is complex and environment-dependent. This study contributes toward the understanding of multimodal interaction in AR and highlights the potential of escape rooms as research platforms, paving the way for future studies to explore other combinations of instruction modalities.

Introduction

Augmented reality (AR) is a technology that overlays digital information, such as images, sounds, and interactive elements, onto the real-world environment, typically through devices like smartphones, tablets, or AR glasses (Hertel et al., 2021). Unlike virtual reality, which creates entirely immersive digital environments, AR enhances the physical world by seamlessly integrating virtual elements, allowing users to interact with both simultaneously.

AR has gained significant relevance in entertainment, being used to create immersive experiences, such as interactive games and storytelling, that engage users in new and innovative ways. For example, AR-based gaming applications like Pokémon GO have revolutionized how players interact with digital content, blending the virtual and real worlds to enhance gameplay (Mendoza-Ramirez et al., 2023).

The integration of AR with interactive experiences, such as escape rooms, is an emerging trend that combines physical and digital elements to create highly immersive environments. By incorporating multimodal interaction, and leveraging gestural, auditory, and touch inputs, these experiences enhance user engagement, accessibility, and the sense of realism, making them more dynamic and inclusive (Rakkolainen et al., 2021). However, despite significant advancements in the field, there is a lack of research comprehensively evaluating both technical and experiential aspects of multimodal systems (Merino et al., 2020).

The purpose of this project is to design, develop and evaluate a novel augmented reality escape room that leverages multiple modalities to create an interactive, and immersive user experience. Unity was chosen as a development platform due to its extensive AR toolkits. The escape room was designed for deployment on an iPad Pro to provide a more accessible experience compared to games designed for headsets (Mendoza-Ramirez et al., 2023). Moreover, the iPad Pro features a LiDAR scanner that alleviates markerless object positioning in AR (Cao et al., 2023). Leveraging the built-in capabilities of the device, the escape room incorporates the input modalities of touch, speech and movement with the system's visual output modality.

Although there are several existing AR implementations of escape rooms (Wild et al., 2021; Plecher et al., 2020; Zeng et al., 2020), how the AR environment and choice of modalities may impact interaction compared to a traditional escape room has yet to be thoroughly explored. Notably, an essential component of escape room puzzle design is the instructions that are provided to the player, which can significantly impact user experience and performance (Clarke et al., 2017). Hence, this study examines the impact of speech versus textual instructions on user performance and experience.

Zhao et al. (2024) conducted a similar comparison of the effect of auditory and visual instructions in an AR environment on task completion in a space station. They found visual instructions were more effective, suggesting that visual instructions provide information more intuitively and avoid interference from background noise. However, in the context of an AR escape room where the visual modality is already heavily burdened and instructions are less

complex, auditory instructions may instead enhance immersion (Wild et al., 2021), reduce cognitive load (Wickens, 1976) and lead to improved task performance (Barron, 2004).

Consequently, the study aims to respond to the following research question:

How do auditory and visual instruction modalities in an AR escape room affect user task performance and user experience?

Background

AR and multimodal interaction

This section contains literature related to the field of augmented reality and multimodal interaction. This includes AR frameworks, state-of-the-art implementations and methods of realising multimodal interaction. The literature provides insight into the strengths and weaknesses of our chosen implementation by placing it in a larger context.

Mendoza-Ramirez et al. (2023) offer an overview of the development, applications and challenges of AR. The authors propose a classification of AR technologies based on whether they require visual triggers for object placement or not. Markerless AR allows for greater flexibility and immersion than marker-based AR, being commonly used in video games. However, a limitation of this technology is that object placement may not be as accurate or reliable as in marker-based systems. For example, object recognition and tracking can be negatively impacted by adverse lighting conditions, occlusions and camera quality.

In addition, Mendoza-Ramirez et al. (2023) discuss devices that can be used to enable AR applications, including smartphones. An advantage of building AR experiences for smartphones is that the experience becomes available to a large user group, which makes funding, developing and evaluating an actual AR escape room more feasible. On the other hand, the experience is less immersive than it would have been using a headset, as smartphones have relatively small display sizes and limited interaction modalities. Cao et al. (2023) delve deeper into the field of mobile augmented reality (MAR), providing guidelines for designing, building and evaluating applications. The authors emphasise the importance of including usability-related metrics, such as user perception of information, manipulation and task-oriented outcome.

Hertel et al. (2021) propose a comprehensive taxonomy for interaction techniques in AR, discussing the advantages and limitations of each method. Moreover, the authors outline general trends within AR, such as a shift toward multimodality and increasingly subtle interactions. Similarly, Rakkolainen et al. (2021) present how different modalities can be used and combined in extended reality, including speech, sound, and haptics. The authors highlight the combination of auditory and visual modalities as particularly common in AR, improving immersion and accessibility.

AR escape rooms

This section discusses literature on escape rooms in extended reality, shedding light on existing implementations and acting as a guideline for what is feasible and best practice. Moreover, it highlights research gaps and potential for innovation within AR escape rooms.

Kleinman et al. (2024) discuss the advantages of using escape rooms for research. The authors argue that escape rooms provide a relatively natural and familiar environment for evaluation, while also acting as a behavioural sandbox where data easily and non-intrusively can be collected. However, existing implementations of AR escape rooms often focus on the technical implementation, without thoroughly evaluating how the user experience is impacted by the mixed reality environment and the different interaction modalities. For instance, Zeng et al. (2021) present an AR escape room implemented using a deep convolutional neural network and deployed on the Microsoft Hololens, exploring the effectiveness of object recognition algorithms. The escape room includes NPCs that the player can interact with, an innovative feature enabled by the virtual elements of the experience.

Another example of a recent implementation is provided by Plecher et al. (2020), who developed an AR escape room with a competitive and cooperative playing mode. The escape room accommodates multiple players, introducing a social dimension to the experience. The game is deployed on a handheld device and takes approximately 50 minutes to complete, including a broad variety of puzzles. The authors discuss the opportunities and limitations of AR escape rooms compared to physical, board game and video game escape rooms. Although the AR escape room was appreciated, participants generally preferred the traditional experience, expressing that it was tedious to interact with the handheld device for so long.

A project with a larger emphasis on user experience is Unbody, presented by Wild et al. (2021). Unbody is an escape room with a poetry theme, built with Unity and deployed on the Hololens. The evaluation of Unbody includes the system usability scale (SUS) and the spatial interaction evaluation (SPINE). Using both in combination provides a comprehensive understanding of how system design relates to different aspects of usability in the AR environment. The authors note the importance of multimodality in the experience, emphasising how audio can affect immersion and engagement.

The implementation of Wild et al.'s (2021) escape room is based on the escapED framework, which deconstructs the design into six aspects which should be defined in sequential order: participants, objectives, theme, puzzles, equipment and evaluation (Clarke et al., 2017). Firstly, it is important for developers to understand their participants. This involves defining the user type, length of experience, difficulty, mode (cooperation vs competitive) and scale (how many users the experience can accommodate). For instance, a longer escape room could be unpleasant for people prone to motion sickness and greatly increases the time required to evaluate the system.

Next, the developers must consider the objectives of the game, including learning and behavioural change objectives, whether the game should be solo- or multi-disciplinary, which soft skills it should develop and which problem-solving challenges it should contain (Clarke

et al., 2017). Considering the skills involved in solving the process is an important step in determining the difficulty of the game. A too difficult game risks excluding certain groups of players, whereas a too easy game risks providing an unsatisfying user experience.

The theme of the escape room covers the method of escape and narrative design as well as contributing aesthetic elements (Clarke et al., 2017). Designing the puzzles requires considering how the puzzles relate to the learning objectives and how the player should receive instructions and hints. As for equipment, this includes location and space design, physical and technical props and actors. Traditionally, escape rooms require an extensive amount of equipment and a designated space, but AR escape rooms can enable a prop-less, location-independent experience (Cao et al., 2023).

Finally, evaluating the system necessitates testing the game experience, gathering information about participants' subjective experiences, evaluating the learning objectives, adjusting the system and if necessary, resetting puzzles (Clarke et al., 2017). It is emphasized that this is an iterative process and that the puzzles may need to be adjusted to be appropriate for different user groups.

Evaluation

This section presents literature related specifically to the evaluation of the AR escape room. It introduces different frameworks and methods that can be used to evaluate specific aspects of the user experience, as well as their strengths and weaknesses.

Merino et al. (2020) discuss evaluation methods in XR, identifying two primary approaches: one focused on technical aspects of the implementation and one focused on user experience aspects. Although both approaches are important, there is a lack of research combining both dimensions. Merino et al. also identify a gap in research on evaluating multimodal interfaces. The authors encourage researchers to utilize a breadth of cognitive methods for evaluation, including quantitative methods such as EEG and eye-, head- and body-tracking. They criticise studies that solely rely on qualitative measures since these are likely to be affected by bias (for instance, participants may not want to express negative emotions for fear of offending the researchers).

Guo et al. (2023) discuss the strengths and limitations of various evaluation methods for VR systems. Specifically, the authors encourage integrating the experience into the system to save time and improve user immersion, which can be done through screen recording. They discuss challenges in applying heuristics to modern systems, as they may not accurately capture all aspects of the environment and interactions, which are rapidly changing and developing. A potential solution to this issue is to combine heuristics with other usability models that cover more detailed aspects of the experience. For this project, the spatial interaction evaluation (SPINE) questionnaire (Wild et al., 2020) was chosen for evaluation, providing an in-depth understanding of different aspects of user experience, including the input and output modalities.

Multimodal information processing

This section contains literature focused on theories within psychology and human-computer interaction related to information processing and multimodal systems. Additionally, the section covers literature on the effect of auditory and visual instructions on task performance and user experience. This literature motivates the decision to compare the chosen modalities in the context of an AR escape room while highlighting their strengths and weaknesses and discussing previous research approaches.

The decision to use auditory and visual instruction modalities is grounded in multiple resource theory (Wickens, 1976) and the working memory model (Baddeley & Hitch, 1974). According to multiple resource theory, humans have separate cognitive resources for processing visual, auditory, spatial and verbal information (Wickens, 1976). Tasks that require the same type of resource can lead to cognitive overload. Hence, it is often beneficial to distribute tasks across multiple modalities. Auditory and visual modalities are often complementary, allowing for efficient information processing (Dumas et al., 2009). The working memory model (Baddeley & Hitch, 1974) supports this by proposing separate systems for processing auditory-verbal information and visual-spatial information. In the context of an AR escape room, the visual modality is already engaged when searching for and interacting with puzzles. Consequently, text instructions could risk competing with limited visual resources. Speech instructions leverage the phonological loop in working memory, allowing users to process auditory-verbal information without overloading the visuospatial sketchpad. This could enhance task performance and lead to less cognitive strain.

Speech can be particularly advantageous when the available visual display space is limited, for instance, if the device is an iPad or mobile phone (Sears & Jacko, 2007). Additionally, it does not require users to divert their visual attention from the task. Research indicates that audio can significantly enhance the immersion of the environment, being even more important for user experience than visual aspects of the environment (Wild et al., 2021). On the other hand, speech instructions may be less effective in noisy environments and overload the working memory if the instructions are lengthy or complex, leading to information loss (Barron, 2004). Text has the advantage of persistence, enabling users to digest the instructions at their own pace. This might be preferable in a high-stress environment such as an escape room.

Zhao et al. (2024) compare visual and auditory instructions in the context of a space station. Their results indicate that visual instructions are superior, leading to a lower cognitive load and improved task performance for the participants. The authors propose that this may be due to auditory instructions not being sufficiently precise for the complex tasks, and background noise causing interference and information loss. This suggests that the effect of the instruction modality is largely dependent on the environment. For less complex tasks completed in a more quiet and controlled environment, it is possible that audio could surpass text as an instruction modality. In an AR escape room, speech instructions could provide increased immersion and engagement (Wild et al., 2021), increasing participant focus and motivation.

Methods

Experiment structure

The experiment was constructed as a between-study design where each participant group played a different version of the escape room game. The two groups, A and B, received in-game instructions for each puzzle through different modalities. For group A, all instructions were given in a visual format, in this case, text instructions displayed next to the task. For group B, all instructions were given in an auditory format, in this case, speech was played when the participant moved close to a task.

The procedure of the experiment was the same with each participant answering a pre-experiment survey, playing the AR escape room game, and answering a post-experiment survey after playing the game.

Participants

Participants were recruited through convenience sampling by the researchers. A recruitment form was distributed across social networks, and individuals near the testing location were invited to join at unbooked time slots. Only English-speaking participants were invited, as this was the language in which the escape room instructions were delivered. Beyond this, no specific participant screening process was implemented. However, relevant demographic data was collected through a pre-experiment survey. Initially, the aim was to recruit a minimum of 20 participants, balancing statistical power with adhering to the study's time and resource constraints. Notably, this is still relatively small for a between-groups study, and may not be enough to detect medium or small effects.

Pre-experiment survey

The pre-experiment survey contained demographic questions concerning age, gender and participants' previous experience with augmented reality on a five-point Likert scale (view Appendix A). Collecting this data was essential for the between-group study design, allowing for the assessment of potential demographic influences and ensuring balanced group comparisons.

Experiment

The experiments were conducted over two days at the KTH campus, a comfortable and accessible place for the majority of participants, many of whom were students. The testing location was a dedicated room where the participant could complete the experiment undisturbed and anonymously, reducing the risk of interference and outside distractions. The room was arranged to provide a large area for participants to walk around and explore the escape room without potentially colliding with real-world objects. Researchers were present throughout all experiments to ensure that the AR environment was set up correctly and to

assist participants with any technical issues that could arise. Participants were randomly assigned to a condition, with the speech instruction experiments running the first day and the textual instructions running the second day.

Before commencing the experiment, participants were informed that they would be tasked with solving puzzles inside an escape room with a time limit. Additionally, the researcher explained the user interface of the game to them, including how to view how many puzzles they had completed and how much time they had left. Following the traditional structure of an escape room (Clarke et al. 2017), participants were left to work out the remaining mechanics of the escape room and individual puzzles on their own. All experiments were conducted individually and all participants wore headphones to reduce the influence of potential background noise.

The following quantitative data was collected from the experiment:

- Completion rate: To what degree the participant was able to complete a certain task
- Whether or not the participant succeeded in completing the escape room within the time limit
- The time left of the time limit upon completing the escape room

In addition to these data points, observational notes were taken for each participant during the experiment. These notes included but were not limited to: participants experiencing difficulties with specific tasks, general behaviour while playing, participant comments, and technical problems that occurred during the session.

Post-experiment survey

The post-experiment survey contained questions regarding the user experience of the escape room (view Appendix B). The survey was largely based on SPINE, an evaluation framework that considers six constructs of the application's usability: system control, navigation, manipulation, selection, and input and output modalities (Wild et al., 2020). This framework was deemed suitable for the evaluation as it places a large significance on the system's combination of modalities and how they affect the user experience, which closely aligns with the study's research question.

In addition to SPINE, the survey contained questions regarding the difficulty and enjoyment of the three individual puzzles, as well as an open-ended question where participants could write freely about their experience. The data gathered from this question was used in combination with the observational notes taken during the experiment to form the basis for a qualitative evaluation of the system.

AR environment

A demo of the escape room is available at the following link:
https://youtu.be/E4CDhO39AbM

The augmented reality environment was built using the Unity (2022.3.46f1) game engine. An iPad was used to play the game and thus the project was built primarily for IOS. The development of the escape room was informed by the guidelines proposed by Clarke et al. (2017). The escape room had a jungle aesthetic where each puzzle was represented by an animal. Each animal used a specific modality to solve its task. These animals were custom-made using DCC software and personalities were given to each for increased immersion for the participant. They consisted of:

- *Luna the Cat*: A game where the task was to pat a cat's head and rub its belly through the use of touching and swiping on the iPad screen.
- *Kaa the Cobra*: A game where the task for the participant was to match a cobra's movement with the iPad, by moving it horizontally side to side.
- *Peri the Parrot*: A game where the task for the participant was to give an answer to a riddle which the parrot told. The answer was given through the use of audio input from the participant.

The animal models can be viewed below (Figure 1). Interaction for the two former games made use of built-in Unity features while the latter third made use of OpenAI speech API. To ensure consistent conditions for all participants, the same room was used throughout the experiment where each task was positioned in the same location, i.e task positions were fixed. Completing the room was timed, with each participant getting five minutes to complete the room. Informed by previous research, the duration of the escape room was kept short to avoid participant fatigue and ease data collection (Plecher et al., 2020). Time was automatically tracked by the game and the time left was written down by researchers after the participants completed the experiment. Integrating the completion time measurement in this way was done in line with the recommendations of Guo et al. (2023), reducing the need for researcher interference during the experiment.

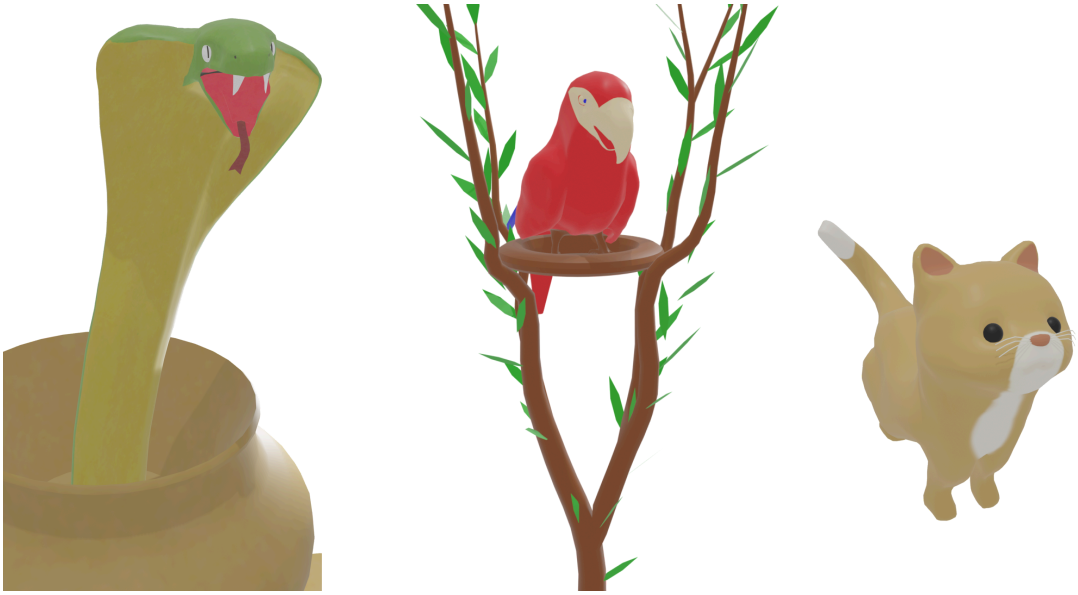


Figure 1. Kaa the Cobra, Peri the Parrot and Luna the Cat.

Data analysis

The quantitative data was statistically analysed by comparing the results of the two groups. Assumptions of normality and variance using Shapiro Wilk's and Levene's test were made. For task performance, completion rates and time left upon completion were compared between groups using a Chi-square test and Mann-Whitney U test, respectively. Differences in user experience according to SPINE and user puzzle ratings were evaluated using t-tests and Mann-Whitney U tests depending on the distribution of the data. The qualitative data is integrated with the quantitative data in the analysis to provide a more holistic understanding of the results.

Results

Participants

A total of 24 (8M, 16F) participants between the ages of 22 to 32 were recruited for the experiment, with 12 participants assigned to each condition. The mean age, AR experience and gender distribution for each group are shown below:

Table 1. Participant demographics.

Group	Mean age	# males	# females	Mean AR experience
All participants	24.50	8	16	2.46
B: Speech instruction	24.033	6	6	2.583
A: Text instruction	24.833	2	10	2.333

Notably, the gender distribution in the speech group was balanced, whereas the gender distribution in the text group was skewed with a higher number of female participants.

Since the speech group data violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.73$, $p = 0.002$), a Mann-Whitney U test was conducted to compare the mean age between groups. There was no significant difference in mean age between the speech ($M = 24.03$, $SD = 2.84$) and text group ($M = 24.83$, $SD = 2.66$), $U = 57.00$, $z = -0.87$, $p = 0.39$.

When comparing mean AR experience between groups, the text group data violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.83$, $p = 0.023$). As a result, a Mann-Whitney U test was conducted. There was no significant difference in mean AR experience between the speech ($M = 2.58$, $SD = 0.996$) and text group ($M = 2.33$, $SD = 1.07$), $U = 86.50$, $z = 0.84$, $p = 0.396$.

Task performance

To assess task performance, the time left after completing the room, completion rates for both the room and puzzles were measured during the experiment. In the speech group, all participants completed the escape room. In comparison, the text group had a room completion rate of 66.67%, with each individual task having the following completion rate:

Table 2. Task completion rates for the text condition.

Task	Luna the Cat	Kaa the Cobra	Peri the Parrot
Completion rate (%)	83.33	66.67	83.33

Differences in room completion rate between the speech and text groups were examined using a Chi-square test. The proportion of participants that completed the room between conditions did not differ significantly ($x^2 = 2.7$, $df = 1$, $p = 0.10$). The effect size was medium, as indicated by Cramer's $V = 0.34$.

The time limit for completing the room was 5 minutes, with the mean completion time being 61.75 seconds (SD = 51.82). Since the data for the text group violated the assumption of normality (W = 0.78, p = 0.005), a Mann-Whitney U test was performed to compare time left between the speech (M = 91.08, SD = 45.95) and text group (M = 32.42, SD = 40.33). There was a statistically significant difference in time left, with the speech group completing the escape room faster, U = 120.00, z = 2.77, p = 0.006. The effect size was large, r = 0.57.

User experience

Data regarding user experience was collected through the SPINE questionnaire (view Appendix B) and participants' puzzle difficulty and enjoyment ratings. Additionally, qualitative data from observations during the experiments and participant comments is included in the analysis.

SPINE constructs

The SPINE framework evaluates system usability through six constructs: system control, manipulation, navigation, selection, and input and output modalities. Items were rated by participants on a 7-point Likert scale. In the context of this study, the input modalities constructs included questions concerning the touch screen, the movement-based interaction, voice input, and the iPad Device. The output modalities construct contained questions concerning the delivery of the instructions, ease of interaction, spatial arrangement, and audio output. The box plot below indicates the distribution of scores across all participants. Item OM4, concerning the audio output, has been excluded in this comparison, as this was only experienced by the speech group.

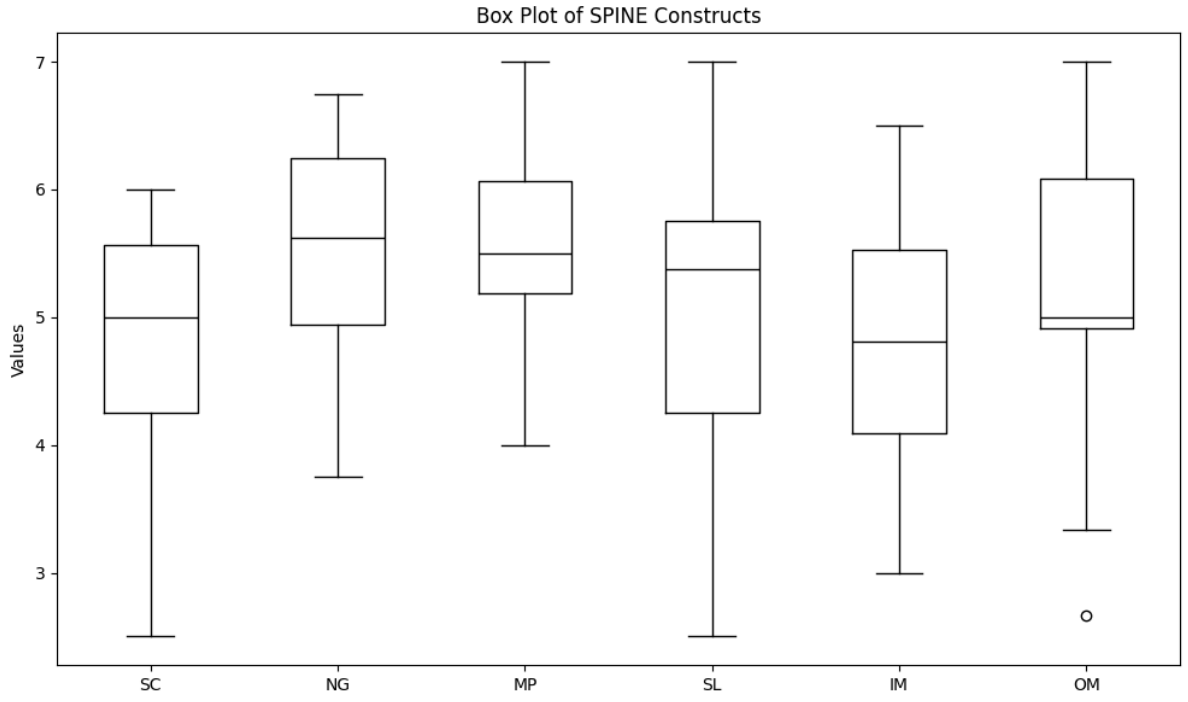


Figure 2. Box plot of SPINE constructs for all participants.

Table 3. Mean and SD of SPINE constructs for all participants.

Construct	**Mean**	**Standard deviation (SD)**
System control (SC)	4.84	0.90
Navigation (NG)	5.52	0.91
Manipulation (MP)	5.55	0.74
Selection (SL)	5.05	1.13
Input modalities (IM)	4.85	0.94
Output modalities (OM)	5.24	1.15

The mean and standard deviation for each construct are summarised above. The scores for the constructs were relatively uniform, with manipulation scoring the highest and system control scoring the lowest. Overall, the scores indicate that the system did not have any major usability issues.

Box plots for each construct for each condition (speech and text) can be viewed in Appendix C. Differences in constructs between the conditions were statistically analysed to determine whether the instruction modality affected user experience.

For system control, the mean value was 4.85 (SD = 0.69) and 4.83 (SD = 1.10) for the speech and text groups respectively. As the text group data violated the assumption of normality ($W = 0.85$, $p = 0.04$), a Mann-Whitney U test was conducted. The test showed no significant difference in system control between conditions, $U = 64.00$, $z = -0.46$, $p = 0.66$. The effect size is negligible, $r = -0.09$.

For navigation, the mean value was 5.88 (SD = 0.93) and 5.17 (SD = 0.77) for the speech and text groups respectively. As the speech group data violated the assumption of normality ($W = 0.77$, $p = 0.004$), a Mann-Whitney U test was conducted. The test showed a significant difference in system control between conditions, $U = 110.00$, $z = 2.19$, $p = 0.03$, with a medium effect size, $r = 0.45$. The text group experienced navigation as more difficult than the speech group. However, considering the number of comparisons made, a Bonferroni correction would suggest that this result is not significant. Therefore, it should be interpreted cautiously.

For manipulation, the mean value was 5.42 (SD = 0.84) and 5.69 (SD = 0.63) for the speech and text groups respectively. A student's t-test showed no significant difference in system control between conditions, $t(22) = -0.89$, $p = 0.38$. The effect size is small, Cohen's $d = -0.36$.

For selection, the mean value was 4.94 (SD = 1.12) and 5.17 (SD = 1.18) for the speech and text groups respectively. A student's t-test showed no significant difference in system control between conditions, t(22) = -0.49, p = 0.63. The effect size is small, Cohen's d = -0.20.

For the input modalities, the mean value was 4.97 (SD = 1.11) and 4.73 (SD = 0.76) for the speech and text groups respectively. A student's t-test showed no significant difference in system control between conditions, t(22) = 0.62, p = 0.54. The effect size is small, Cohen's d = 0.25.

For the output modalities, the mean value was 5.44 (SD = 0.96) and 5.03 (SD = 1.32) for the speech and text groups respectively. A student's t-test showed no significant difference in system control between conditions, t(22) = 0.88, p = 0.39. The effect size is small, Cohen's d = 0.36.

Puzzle ratings

Puzzle difficulty was rated by participants on a five-point Likert scale.

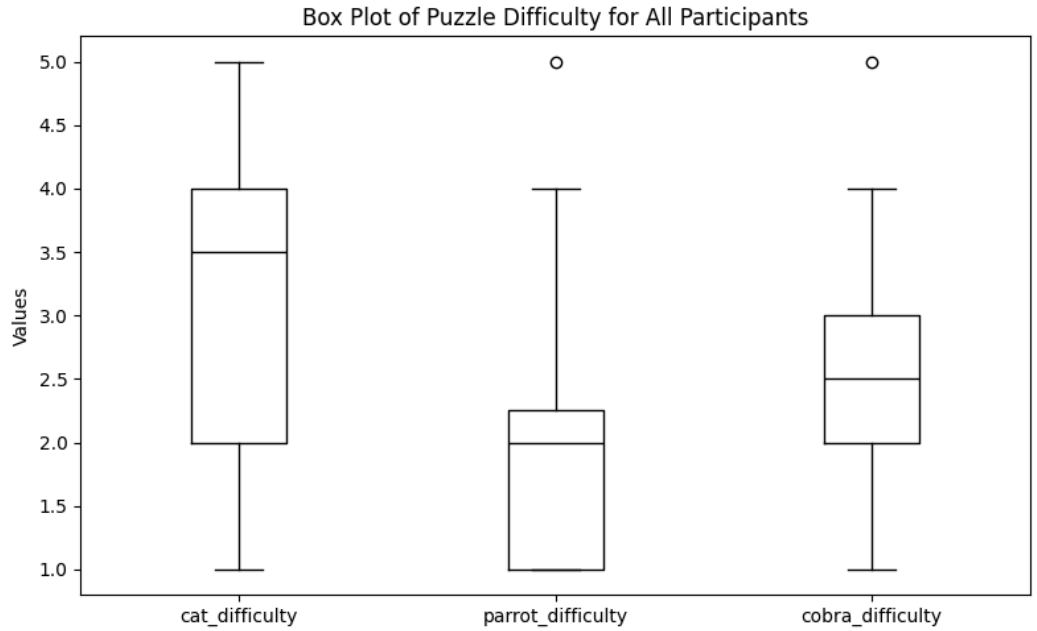


Figure 3. Puzzle difficulty ratings for all participants.

Table 4. Participant ratings of puzzle difficulty.

Puzzle	**Mean**	**Standard deviation (SD)**
Luna the Cat	3.04	1.33
Peri the Parrot	2.00	1.14
Kaa the Cobra	2.54	1.14

The cat puzzle was rated the most difficult and the parrot puzzle as the least difficult. Notably, many participants misunderstood the intended interaction with the cat, attempting to pat it by reaching out toward it instead of touching the screen.

Qualitative feedback from participants during the experiments and in the survey suggests that the textual instructions could make completing the puzzles more challenging, as this required them to position themselves correctly to read the text, distracting them from the task at hand. Another participant expressed that they were too focused on the animal character to notice the less visually engaging text nearby. Conversely, audio instructions could also contribute to the difficulty. One participant missed the instruction for a puzzle, requiring the researchers to intervene and repeat it.

Statistical tests were conducted to examine whether instruction modality impacted perceived puzzle difficulty. In the speech group, the mean difficulty rating for the cat puzzle was 3.08 ($SD = 1.31$), whereas in the text group, the mean rating was 3.00 ($SD = 1.41$). Since the cat difficulty ratings for the text group violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.89$, $p = 0.04$, a Mann-Whitney U test was conducted to analyze whether the instruction modality impacted the perceived difficulty of the cat puzzle. There was no significant difference in perceived difficulty between the speech and text group, $U = 74.00$, $z = 0.12$, $p = 0.93$. The effect size is negligible, $r = 0.02$.

The mean difficulty rating for the parrot puzzle was 1.83 ($SD = 1.03$) for the speech group and 2.17 ($SD = 1.27$) for the text group. Both the speech and text group data violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.80$, $p = 0.01$ and $W = 0.82$, $p = 0.02$ respectively). As a result, a Mann-Whitney U test was conducted. There was no significant difference in perceived difficulty for the parrot puzzle between the speech and text group, $U = 61.00$, $z = -0.64$, $p = 0.52$. The effect size is small, $r = 0.13$.

The mean difficulty rating for the cobra puzzle was 2.58 ($SD = 0.90$) for the speech group and 2.50 ($SD = 1.38$) for the text group. A student's t-test was conducted, indicating no significant difference in perceived difficulty between the speech and text group, $t(22) = 0.18$, $p = 0.86$. The effect size is negligible, Cohen's $d = 0.07$.

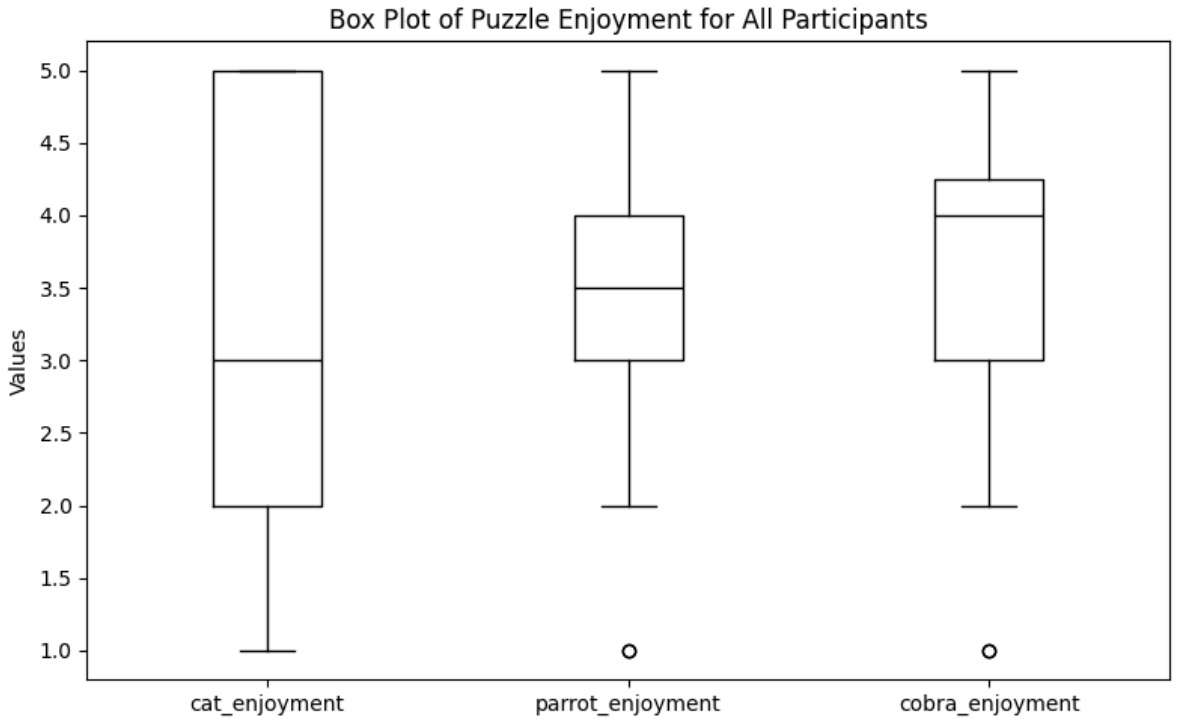


Figure 4. Box plot of enjoyment ratings for all participants.

Table 5. Participant ratings of puzzle enjoyment.

Puzzle	**Mean**	**Standard deviation (SD)**
Luna the Cat	3.38	1.28
Peri the Parrot	3.38	1.17
Kaa the Cobra	3.75	1.15

Participant enjoyment was relatively uniform across puzzles, with many participants commenting that it was a fun experience in the post-experiment survey. One participant in the speech group expressed that it was particularly enjoyable to be able to talk to the animals.

In the speech group, the mean enjoyment rating for the cat puzzle was 3.33 (SD = 1.07). In the text group, the mean rating was 3.42 (SD = 1.51). Since the ratings for the text group violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.83$, $p = 0.02$), a Mann-Whitney U test was conducted to compare whether the instruction modality impacted the perceived difficulty of the cat puzzle. There was no significant difference in perceived difficulty between the speech and text group, $U = 69.50$, $z = -0.14$, $p = 0.90$. The effect size is negligible, $r = 0.03$.

The mean enjoyment rating for the parrot puzzle was 3.50 (SD = 1.24) for the speech group and 3.25 (SD = 1.14) for the text group. A student's t-test was conducted, indicating no

significant difference in enjoyment between the speech and text group, $t(22) = 0.51$, $p = 0.61$. The effect size is small, Cohen's $d = 0.21$.

The mean enjoyment rating for the cobra puzzle was 3.67 (SD = 0.78) for the speech group and 3.83 (SD = 1.47) for the text group. Both the speech and text group data violated the assumption of normality based on Shapiro-Wilk's test ($W = 0.84$, $p = 0.03$ and $W = 0.76$, $p = 0.003$ respectively). As a result, a Mann-Whitney U test was conducted, indicating no significant difference in enjoyment between groups, $U = 54.00$, $z = -1.04$, $p = 0.28$. The effect size is small, $r = 0.21$.

Discussion

While the experiment yielded interesting results, some limitations were present. These limitations were related to technical implementations and design elements in relation to human-computer interaction principles.

Technical limitations

Certain technical limitations existed in each task which may have implications for the experiment. These were mainly related to the Unity implementation or task design, and could likely have been resolved given more time. For Peri the Parrot, the main limitation involved the speed of input and output to the speech API. There was often a delay between the user's answer and the system's response, causing some players to wrongly believe that their answer was incorrect. The importance of system feedback for speech input is underscored by Rakkolainen et al. (2021), who warn that inefficient error handling can lead to a slow and tedious interaction. Similarly, Kaa the Cobra could have made use of more intuitive gesture recognition. Currently, the task only registers horizontal movement and not any rotation of the iPad. Utilizing both movement and rotation could have yielded a more intuitive experience. The task of Luna the Cat could be made more intuitive by implementing gesture-based input as opposed to touch input. The instructions could also have been adjusted to minimize misinterpretations.

Participants

The participants in the study were aged between 22 and 32 years, resulting in findings that may only be representative of a younger demographic. Cognitive abilities can vary significantly with age, with older adults often experiencing greater difficulties in processing multiple channels of information simultaneously (Barron, 2004). This decline in information processing resources may inhibit their ability to navigate complex multimodal stimuli. Additionally, existing literature indicates that AR interfaces can pose a challenge for older adults, due to the demographic's lower digital literacy and increased physical and cognitive limitations (Seifert & Schlomann, 2021). Similarly, children still developing their information processing abilities generally have lower speech comprehension (Barron, 2004) and reading

comprehension than young adults. Consequently, the results of this study may not be generalisable to other age groups.

Additionally, the level of familiarity with augmented reality may be higher among the participants than in the average population. The mean level of experience with AR across all participants was 2.46 on a five-point Likert scale, with 5 representing “Highly experienced” and 1 representing “No experience”. Although AR received widespread market adoption with the release of Pokemon Go and facial filters on Snapchat and Instagram (Cao et al., 2023), it is still a relatively new technology. Since a large proportion of the participants were students, they are perhaps more likely to have come into contact with AR technology as part of their studies and may have greater technological literacy than the average individual.

No significant differences in age or AR experience were found between the two conditions, which strengthens the credibility of the results, indicating that it is unlikely that these demographic differences impacted the results. However, the gender distribution of the text group had an overrepresentation of female participants. Although AR experience, age and technical literacy likely have a larger impact on task performance and user experience than gender, a more balanced distribution across both groups would have been preferable.

Task performance

The results indicate that speech instructions are better for task performance than text instructions in an AR escape room context, contradicting the findings of Zhao et al.’ study exploring task performance in a space station (2024). However, this can likely be explained by the differences in environment, with the escape room having fewer disturbances in the form of background noise and less complex instructions for the tasks. Rakkolainen et al. (2021) state that the speech modality is generally well-suited for AR settings, where the user’s visual attention is already occupied and rendering text can be challenging. Additionally, users often wear headphones that isolate them from outside noise, easing the auditory information delivery. This underscores the importance of considering the setting when deciding upon modalities in a system, especially in multimodal environments such as augmented reality. It is important to assess how the strengths and weaknesses may be appropriate for different kinds of tasks, and whether the chosen modalities complement each other or lead to interference (Wickens et al., 1976).

User experience

When measuring user experience through SPINE, no significant differences between the speech and text conditions were found, except for in the navigation construct. Though this result is not significant after applying a correction, it is possible that it could have retained significance with a larger sample size. The navigation construct contained questions regarding the user’s ease of finding and interacting with puzzles, as well as positioning themselves correctly in the physical space. Notably, some participants commented that

reading the text instructions distracted them from the task since it required them to step back from the animal and view them from an appropriate distance. On the other hand, participants in the speech condition did not need to be in a particular position to hear the audio instructions, beyond being close to the animal. This could explain why this construct showed a difference between groups. In addition to affecting user experience, this difference can clarify why the completion time was lower for the speech group compared to the text group.

Overall, the results from the SPINE questionnaire indicate that the AR escape room does not have any major usability issues, indicating that it served as an appropriate and sufficiently well-developed environment for evaluating the effect of the instruction modalities. This supports Kleinman et al.'s (2024) view that escape rooms are a good environment for research, enabling rich data collection while providing the participant with a fun, immersive experience. The markerless design of the escape room allows for easy generalization to other environments (Mendoza-Ramirez et al., 2023). Since it is built for smartphone and tablet deployment, the game is highly scalable and could be further developed into a widely accessible product (Cao et al., 2023).

Nevertheless, there is substantial room for improvement, with the existing implementation primarily serving as a proof of concept. Increased instructions and hints could improve the experience for participants (Clarke et al., 2017). Additionally, the experience could be made more gamified and diverse, with different puzzles being randomly generated each time. Including more complex interactions and more modalities would likely improve user experience (Wild et al., 2021; Rakkolainen et al., 2021), in line with the trend towards increased multimodality in AR applications (Hertel et al., 2021).

Puzzle evaluation

Participants' puzzle ratings indicate that the experience was fun, balanced and appropriately challenging for most participants. There were no significant differences in enjoyment and difficulty of the puzzles between conditions, nor any large differences in enjoyment and difficulty between puzzles. Peri the Parrot task was rated as the easiest task, while Luna the Cat was the most difficult task according to the same metrics. Interestingly, these tasks had the same completion rate in the text group, 83.33%, whereas Kaa the Cobra had a lower completion rate, 66.67%. The higher perceived difficulty of the cat puzzle is likely due to many participants misinterpreting the intended input, attempting to use camera-based gestures rather than the touch screen. The completion rates of the puzzles could have been affected by the puzzle's positions in the escape room, with many users choosing to go to the leftmost puzzle (Luna the Cat) first, then the middle puzzle (Kaa the Cobra), and finally the rightmost puzzle (Peri the Parrot).

Compared to a traditional escape room, the puzzles were relatively simple (Clarke et al., 2017). However, existing research indicates that shorter escape rooms may be preferable when the game requires a handheld device (Plecher et al., 2020). Nevertheless, adding a narrative element to the game and more thematic elements could have elevated the

experience. Zeng et al.'s (2021) AR escape room implementation highlights the opportunities of including virtual NPCs, such as Luna the Cat, Peri the parrot and Kaa the Cobra. The NPCs can effectively replace human actors while offering new avenues for interaction, being able to easily appear, disappear, transform and move around. Through the usage of LLMs, it's even possible to enable natural and dynamic conversations between the player and the NPCs. Given more time, it would have been interesting to build more advanced puzzles and explore further how the AR environment can enable new forms of interaction.

Future directions

For future research, it would be interesting to investigate other combinations of instruction modalities. For instance, displaying virtual silhouettes that demonstrate intended interactions could be another form of visual instruction, with unique strengths and limitations compared to text. Additionally, redundant combinations of modalities could be explored, such as combining visual and auditory instructions. Another direction would be to explore the use of more subtle instruction modalities such as haptics, which are increasingly common in AR applications (Hertel et al. 2021; Rakkolainen et al., 2021). Haptics could provide non-intrusive guidance for players, indicating when they are performing an action correctly or close to reaching a solution. Modern smartphones and tablets have quite advanced multimodal functionality (Cao et al., 2023), enabling a wide range of innovative interaction mechanisms, although additional hardware may be required for certain implementations.

More advanced interaction would be exciting to explore in future iterations of the escape room. We considered incorporating eye tracking into our system, but this was not feasible due to the front-facing and back-facing camera of the iPad not being able to be simultaneously active. However, incorporating hand-tracking to allow players to pet Luna the Cat in the physical space in which the model is displayed is likely possible with current technology. This would better align with player expectations for the interaction. Despite any existing limitations, opportunities for multimodal interaction are likely to steadily increase as the technology continues to mature and grow in popularity (Rakkolainen et al., 2021).

Merino et al. (2020) strongly encourage the usage of physiological measures for evaluation, highlighting their reduced susceptibility to bias. To explore how task performance is affected by instruction modalities in greater depth, it could be beneficial to incorporate other evaluation metrics such as eye-tracking and heart rate. Combined with existing theories in HCI and information processing such as multiple resource theory (Wickens, 1976) and the working memory model (Baddeley & Hitch, 1974), this could provide greater insight into multimodal information processing. Additionally, future research into AR escape rooms should aim to enhance the integration of evaluation measures in the game (Kleinman et al., 2024; Guo et al., 2023). This can be achieved by requiring participants to evaluate puzzles upon completion in-game, enabling seamless data collection that does not interfere with the participant's experience. This approach not only preserves the player's immersion of the gameplay but also provides valuable insight into the player's interactions and preferences.

Conclusion

This study explored the effect of instruction modality on task performance and user experience in an augmented reality escape room. The results indicate that speech instructions offer advantages over text instructions, with participants completing tasks more quickly when guided by speech. This could be due to the speech modality not competing over visual resources that are already engaged when searching for and interacting with the puzzles. Despite this, user experience did not vary between groups except in terms of navigation, with the speech group more easily finding and interacting with puzzles and better understanding how to position themselves in relation to virtual elements. This difference can likely be explained by participants in the text condition having to position themselves correctly to read the text, a limitation not faced by the speech group.

Overall, user experience results indicate that the escape room was an enjoyable and suitable environment for investigating instruction modalities. Ratings for enjoyment and difficulty were similar across the three puzzles, with most participants successfully escaping the room. Future research could explore different user demographics and more complex tasks while incorporating improvements in usability.

References

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47–89). New York. Academic Press.

Barron, A. E. (2004). Auditory Instruction. *In Handbook of Research on Educational Communications and Technology* (pp. 949-978). Routledge.

Cao, J., Lam, K., Lee, L., Liu, X., Hui, P., & Su, X. (2023) Mobile Augmented Reality: User Interfaces, Frameworks and Intelligence. ACM Computing Surveys, 55(9), 1-36. https://doi.org/10.1145/3557999

Clarke, S. Peel, D., Arnab, S., Morini, L., Keegan, H., & Wood, O. (2017). EscapED: A Framework for Creating Educational Escape Rooms and Interactive Games to For Higher/Further Education. *International Journal of Serious Games, 4*(3), 73-86. https://doi.org/10.17083/ijsg.v4i3.180

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In Denis Lalanne, Jürg Kohlas (Eds.) *Human Machine Interaction*, LNCS 5440, Springer-Verlag, Berlin/Heidelberg, pp. 3-26.

Guo, X., Nerella, K. K., Dong, J., Quian, Z., Chen, Y. (2023). Understanding Pitfalls and Opportunities of Applying Heuristic Evaluation Methods to VR Training Systems: An Empirical Study. *International Journal of Human-Computer Interaction, 40*(9), 2168-2184. https://doi.org/10.1080/10447318.2022.2161238

Hertel, J. Karaosmanoglu, S., Schmidt, S., Bräker, J., Semmann, M., & Steinicke, F. (2021). A Taxonomy of Interaction Techniques for Immersive Augmented Reality based on an Iterative Literature Review. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. https://doi.org/10.1109/ISMAR52148.2021.00060

Kleinman, E., & Harteveld, C. (2024). The Untapped Potential of Escape Rooms as Gamified Research Environments. In *CHI Play Companion 24': Companion Proceedings of the 24 Annual Symposium on Human-Computer Interaction in Play* (pp. 276-278). https://doi.org/10.1145/3665463.3678865

Mendoza-Ramirez, C. E., Tudon-Ramirez, J. C., Félix-Herrán, L. C., Lozoya-Santos, J. J., Vargas-Matrinez, A. (2023). Augmented Reality: Survey. *Applied Sciences, 13*(18). https://doi.org/10.3390/app131810491

Merino, L., Schwarzl, M., Kraus, M., Sedlmair, M., Schmalstieg, D., Weiskopf, D. (2020). Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009-2019). In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 438-451). http://dx.doi.org/10.1109/ISMAR50242.2020.00069

Plecher, D. A., Ludl, M., & Klinker, G. (2020). *Designing an AR-Escape-Room with Competitive and Cooperative Mode*. In B. Weyers, C. Lürig, & D. Zielasko (Eds.), *GI VR/AR*

Workshop 2020: 24–25 September 2020, Trier (pp. xx–xx). Gesellschaft für Informatik e.V. http://dx.doi.org/10.18420/vrar2020_30

Rakkolainen, I., Farooq, A., Kangas, J., Hakulinen, J., Rantala, J., Turunen, M., & Raisamo, R. (2021). Technologies for Multimodal Interaction in Extended Reality - A Scoping Review. *Multimodal Technologies and Interaction, 5*(12). https://doi.org/10.3390/mti5120081

Sears, A., & Jacko, J. A. (2007). *The Human-Computer Interaction Handbook.* CRC Press.

Seifert, A., & Schlomann, A. (2021). The Use of Virtual and Augmented Reality by Older Adults: Potentials and Challenges. *Frontiers of Virtual Reality, 2*. https://doi.org/10.3389/frvir.2021.639718

Wickens, C. D. (1976). The Effects of Divided Attention on Information Processing in Manual Tracking. *Journal of Experimental Psychology: Human Perception and Performance, 2*(1), 1-11. https://psycnet.apa.org/doi/10.1037/0096-1523.2.1.1

Wild, F., Marshall, L., Bernard, J., White, E., & Twycross, J. (2021). Unbody: A Poetry Escape Room in Augmented Reality. *Information, 12*(8). https://doi.org/10.3390/info12080295

Wild, F., Vovk, A., & Guest, W. (2020). Evaluating Augmented Reality. Retrieved 2024-12-01 from https://www.overleaf.com/project/5e4bcb5066a1250001728e80

Zeng, H., He, X., & Pan, H. (2021). Implementation of Escape Room System Based on Augmented Reality Involving Deep Convolutional Neural Network. *Virtual Reality, 25*, 585-596. https://doi.org/10.1007/s10055-020-00476-0

Zhao, Y., Li, Y., Jiang, A., Zhang, H., She, H., & Zhan, W. (2024). Effects of Visual and Auditory Instructions on Space Station Procedural Tasks. *Space: Science & Technology, 4.* https://doi.org/10.34133/space.0130

Appendix A - Pre-experiment survey

The form is available online at the following link: https://forms.gle/4PhxshaoFpQfZvyZ9

Pre-experiment survey

The aim of this study is to evaluate a multimodal augmented reality escape room. Participating takes approximately 30 minutes and involves doing this pre-experiment survey, playing the game, and doing a post-experiment survey about your experience.

We hope you have fun trying to solve the puzzles in the game! If you have any questions or concerns you can contact:

Louisa Hirvonen: lhhi@kth.se

* Indicates required question

1. Participant ID *

2. What is your gender? *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Prefer not to say
- ◯ Other: _____

3. What is your age? *

4. How much experience do you have using augmented reality (AR)? *

Mark only one oval.

	1	2	3	4	5	
No e	◯	◯	◯	◯	◯	Highly experienced

Appendix B - Post-experiment survey

The form is available online at the following link: https://forms.gle/rsgRmRx1f1NZ2Y2u6

Post-experiment survey

Thank you for participating in our study!

This form contains questions regarding your experience playing the Multimodal AR escape room. Do not hesitate to let us know if you have any questions or concerns.

* Indicates required question

1. Participant ID *

System control

2. SC1: I felt in control when playing the game. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

3. SC2: The user interface elements clearly communicated their functions. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

4. SC3: I was able to understand what to do next in the game. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

5. SC4: The user interface blended together with the real world. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Navigation

6. NG1: It was easy to find the targets I could interact with. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

7. NG2: I could successfully access the targets without running into obstacles. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

8. NG3: I had difficulties to identify in which direction I need to face. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

9. NG4: It was clear how I needed to position myself to have the best experience. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Manipulation

10. MP1: I was able to go back in the system and try different actions. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

11. MP2: The user interface interrupted my task performance. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

12. MP3: The interface blocked my interaction with the real world. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

13. MP4: The instruction from the system was not clear. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Selection

14. SL1: The size of the user interface elements was sufficient for me to be able to view and interact with them. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

15. SL2: I understood the functionality I used. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

16. SL3: The system provided timely feedback about what was happening. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

17. SL4: The system provided a clear confirmation of the actions I performed. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Input modalities

18. IM1: The motion-based interaction was intuitive. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

19. IM11: I understood how to move so that the system would recognize my actions. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

20. IM2: The touch screen was easy to use. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

21. IM22: I knew how to touch the screen to interact with the virtual object. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

22. IM3: The voice interaction was effortless. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

23. IM33: I understood which voice commands were recognized by the system. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

24. IM4: The device allows for flexible interaction with the system. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

25. IM44: Using the device was simple. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Output modalities

26. OM1: The instructions were delivered in a way that was clear and easy to understand. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

27. OM2: The virtual elements were easy to interact with. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

28. OM3: The spatial arrangement of the content felt natural and appropriate. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

29. OM4: Audio cues were understandable and loud enough. *

Mark only one oval.

	1	2	3	4	5	6	7	
Stro	◯	◯	◯	◯	◯	◯	◯	Strongly agree

Concluding questions

30. How would you rate the difficulty of the cat puzzle? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very difficult

31. How would you rate your enjoyment of the cat puzzle? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very enjoyable

32. How would you rate the difficulty of the parrot puzzle? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very difficult

33. How would you rate your enjoyment of the parrot puzzle? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very enjoyable

34. How would you rate the difficulty of the cobra puzzle? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very difficult

35. How would you rate your enjoyment of the cobra puzzle? *

Mark only one oval.

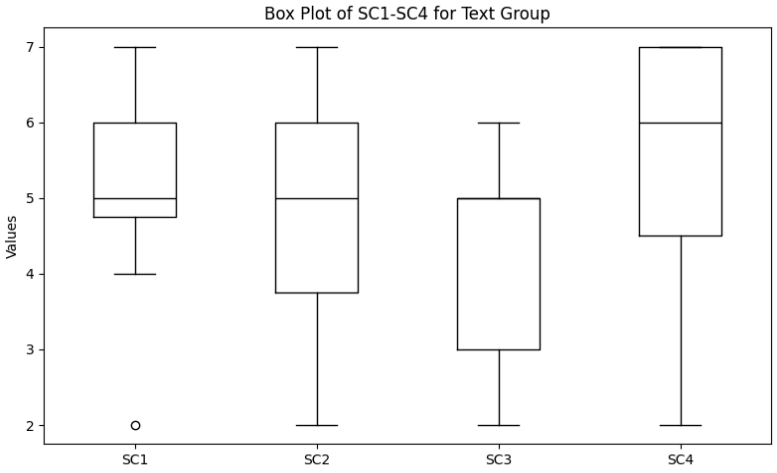
	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very enjoyable

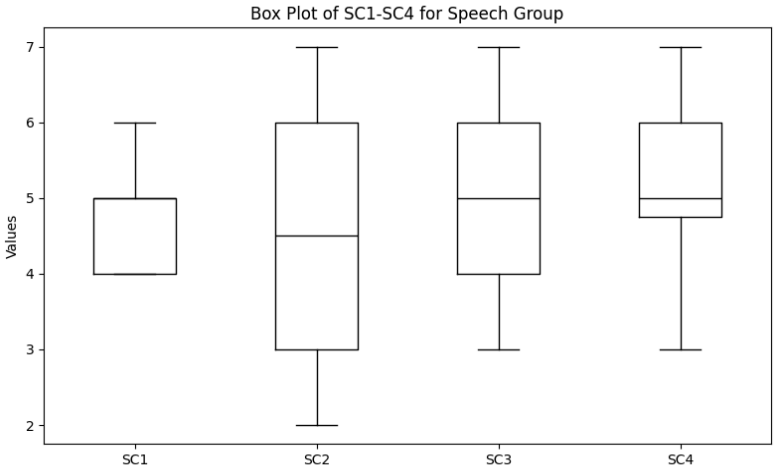
36. Is there anything you would like to add about your experience?

Appendix C - SPINE constructs

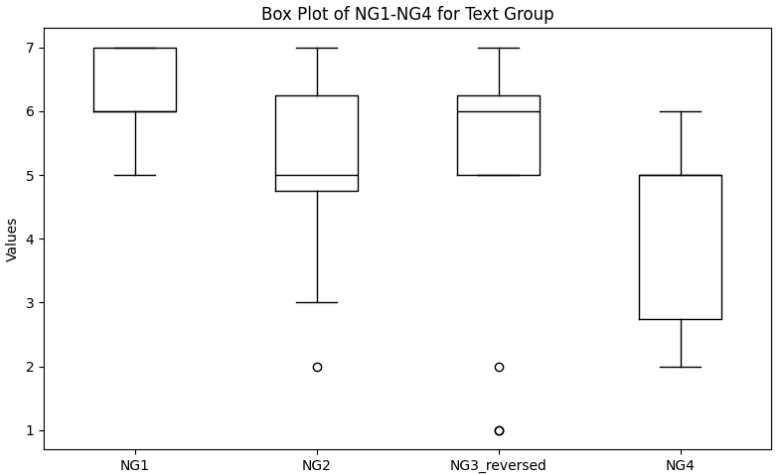
This appendix contains box plots of the six SPINE constructs (system control, navigation, manipulation, selection, and input and output modalities) for the speech and text group. Notably, item four of the output modalities (OM4) has been excluded from the comparison since the measurement is only valid for the speech group, as the text group were not exposed to any auditory output. Additionally, items NG3 and MP2-MP4 have been reversed in polarity due to being negatively coded in the questionnaire.

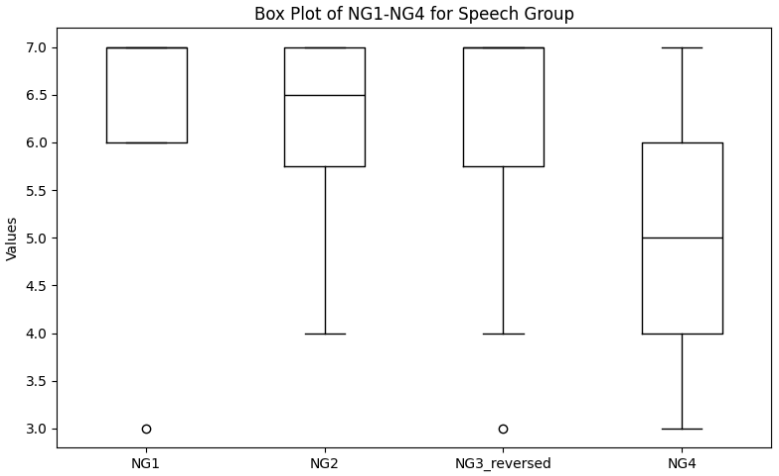
System control



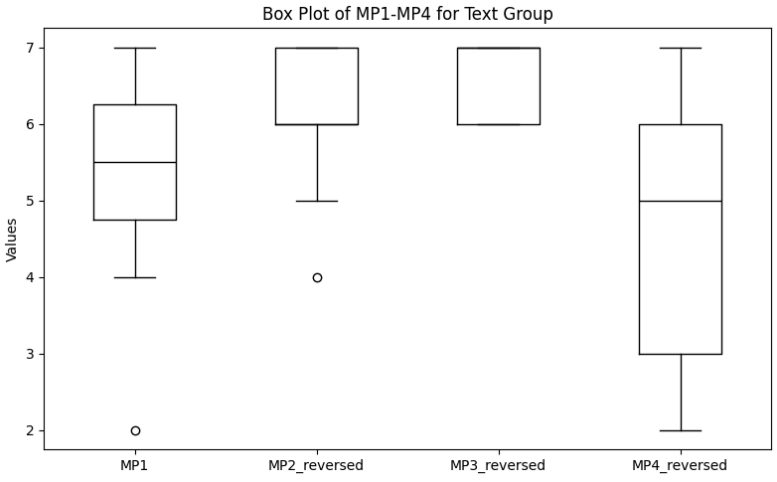


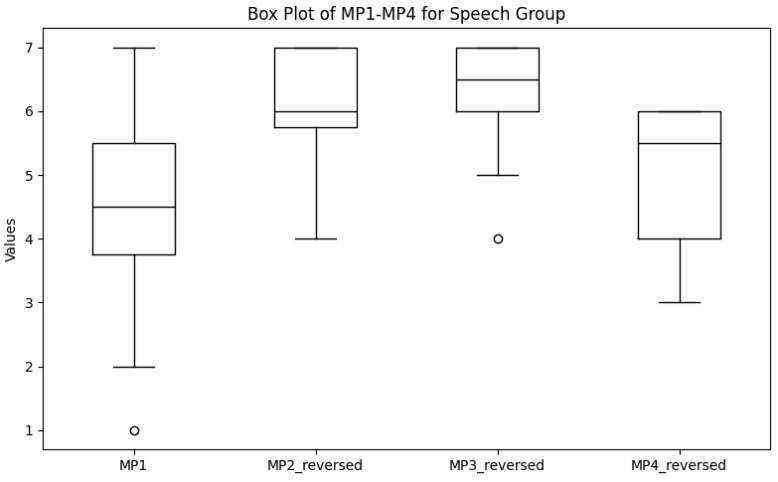
Navigation



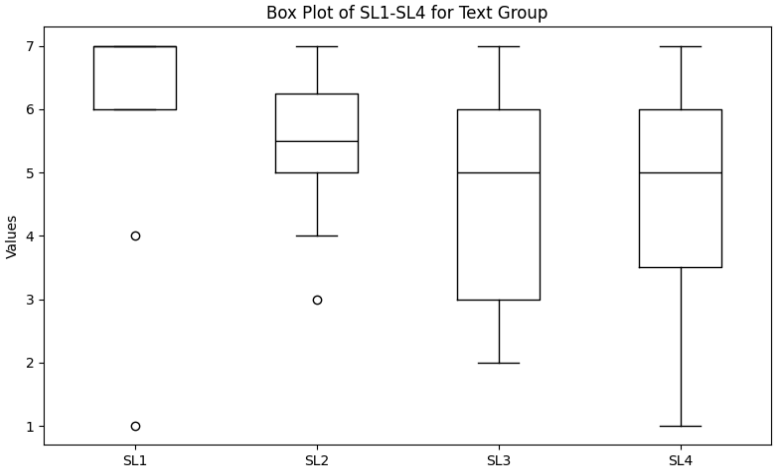


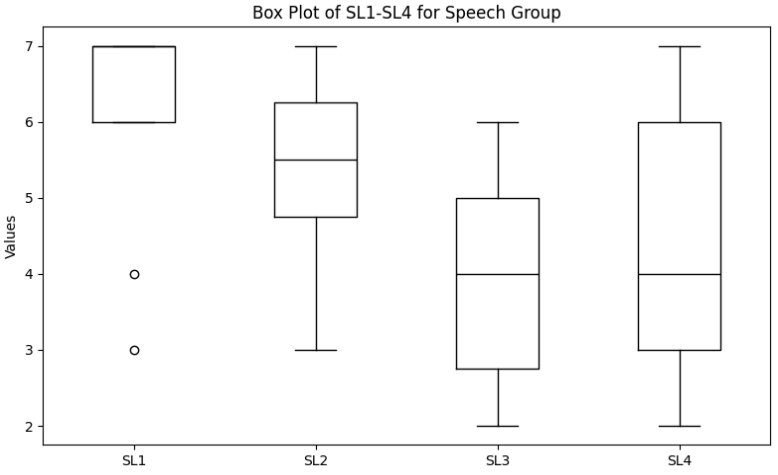
Manipulation



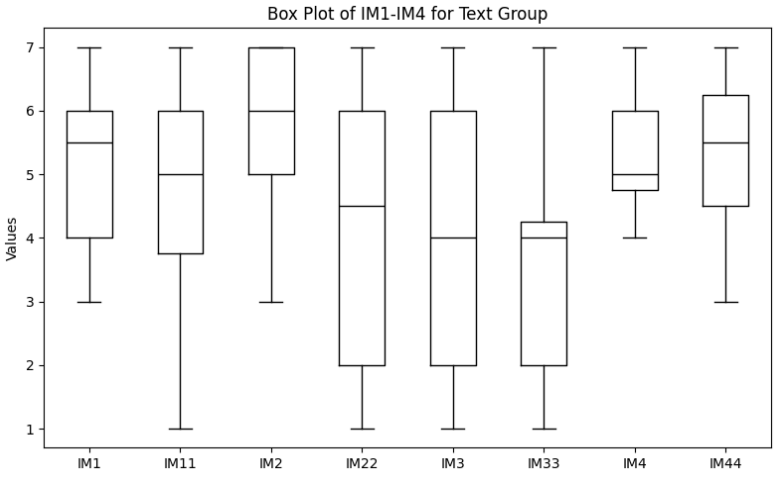


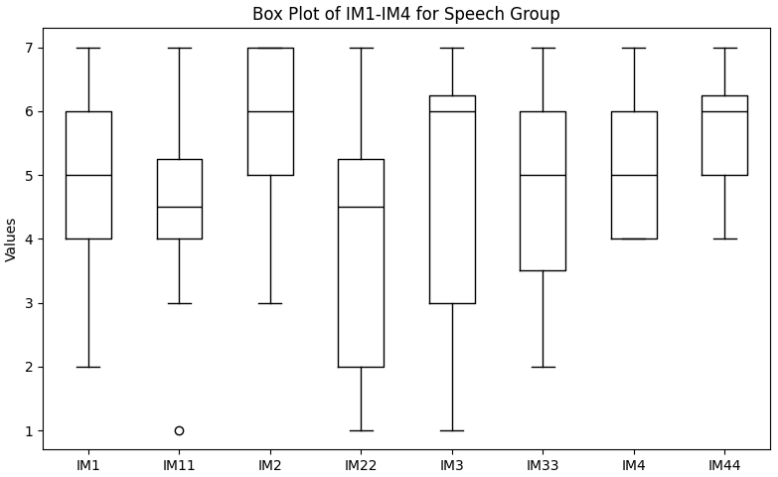
Selection





Input modalities





Output modalities